Implementation and Data Collection

SOCIAL EXPERIMENTATION:
EVALUATING PUBLIC PROGRAMS WITH EXPERIMENTAL METHODS **PART 5**

The government are very keen on amassing statistics. They collect them, add them, raise them to the n*th power, take the cube root, and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases.*

— Sir Josiah Stamp
Inland Revenue Department (England),
1896-1919

This is the fifth in a series of papers on social experimentation. The first four papers focused on the rationale and design of experiments. In this paper, we discuss the more operational issues involved in implementing the experiment in the field and collecting data on the experimental subjects.

In the first part of the paper, we address the following aspects of implementing the experiment:

- Administration of the experimental treatment;
- Gaining the cooperation of program staff;
- Implementing random assignment; and,
- Maintaining the integrity of random assignment.

In the second part of the paper, we discuss the collection of the following types of data that will be required for the experimental analysis:

- Individual identifiers and indicators of treatment status;
- Baseline data;
- Outcome data;
- Program participation data; and,
- Cost data.

Implementation of the Experiment

The early social experiments (*e.g.*, the income maintenance experiments and the health insurance experiment) were administered by the researchers conducting the study. They set up offices in the experimental sites, recruited participants, and issued income maintenance or health insurance benefits according to rules designed especially for the experiment.

These administrative functions required that the experimenters engage in activities far afield from the usual scope of academic research. They were required, for example, to begin by specifying in detail the rules of the experimental program — in the case of the income maintenance experiments, what amounted to a model negative income tax statute, and in the case of the health insurance experiment, a set of detailed health insurance policies. These basic ground rules had to be supplemented with detailed operational forms and procedures to be followed by administrative staff. The researchers then had to set up field organizations to recruit participants — in the case of these two studies, literally by knocking on doors[1] — and administer the experimental benefits. This meant, in effect, running a welfare office or a small health insurance company.[2]

There were certain advantages to administering the experimental treatment directly. It allowed the researchers to specify the experimental program in detail, to ensure that it

[1] In both cases, the experimental sample was recruited by conducting an in-person survey of households in randomly selected dwellings within randomly selected Census tracts in the experimental sites, in order to obtain a sample that was representative of the sites. Those households that were determined to be eligible for the experimental program on the basis of their survey responses were then invited to participate.

[2] The researchers running the health insurance experiment contracted with a commercial claims processing company to handle the experimental insurance claims. Those running the income maintenance experiments, however, directly hired and supervised the staffs administering the experimental payment plans.

was implemented as intended, and to document exactly how it operated.[3]

But there were also distinct disadvantages. It was very costly and time-consuming for researchers unschooled in program administration to develop the requisite expertise for such an undertaking and to create *de novo* the administrative forms and procedures required. Moreover, having done so, it was not clear that the treatment was administered the same way it would have been by regular staff in an ongoing program. In the case of the income maintenance and health insurance experiments, where the experimental benefits were relatively clearly defined in monetary terms, this may not have been a serious problem.[4] In an experiment where the content of the treatment is less well defined — *e.g.*, a counseling program or a drug treatment program — the replicability of the experimental treatment in a regular program could be highly questionable if it is delivered by research staff.

For these reasons, most recent experiments have been administered by the staff of existing programs. For example, tests of employment and training services for welfare recipients have been administered by regular welfare caseworkers, experimental housing vouchers have been administered by the staff of local public housing authorities, and experiments testing home care services for the elderly have been administered by local social service agencies, such as Administrations on Aging. In each case, of course, the experimental treatment was specified, at least in general terms, by the researchers, who oversaw and documented its provision. But in all cases, the researchers were able to take advantage of existing organizational structures, procedures, and staff expertise to deliver an experimental treatment that closely replicated the way an ongoing program would be administered.

There are undoubtedly cases where the experimental treatment is so novel that no existing program structure is appropriate for its administration. (This was, in fact, probably true of the income maintenance experiments.) But in most cases, the policy to be tested in the experiment arises in the context of some existing program and the most sensible way to administer it in the experiment is through that program. And, of course, evaluation of ongoing programs involves administration by regular program staff by definition. In the remainder of this section, then, we assume that the experimental treatment will be administered by regular program staff, under the supervision of the researchers. This means that one of the first tasks in implementing the experiment will be to gain the cooperation of program staff.

Gaining the Cooperation of Program Staff

As noted in the previous paper, program staff tend to resist random assignment for a variety of reasons. Many program operators have a basic *distrust of research and evaluation*. At best, they view it as an irrelevant distraction from their main mission, the day-to-day delivery of services to their clientele. At worst, they see it as a threat to the program. Many view researchers as naive intellectuals who do not understand the reality of program operations, but who nevertheless have the power to adversely affect the funding of their program through an erroneous finding that the program is not working as intended.

This basic distrust is compounded in experimental evaluations by *concerns about random assignment*. Program staff often view random assignment as an unethical denial of service to deserving individuals. Moreover, they expect program applicants who are assigned to the control group to take the same view; as a result, they anticipate having to deal with complaints and objections from the control group, up to and including physical attacks and/or lawsuits. Similarly, they may foresee objections from community groups sympathetic to the program clientele and/or unfavorable publicity in the local media once it becomes generally known that eligible, deserving applicants are being randomly excluded from the program.

Program operators are usually concerned about the *administrative burden and disruption* that a research project, especially one involving random assignment, is likely to entail. The requirement that additional applicants be processed in order to provide for a control group is seen (quite correctly) as creating additional work for a staff that may already be stretched thin and/or raising the costs of a program that is (in the staff's view) already underfunded. In addition, staff must spend time explaining random assignment to applicants and dealing with their questions and complaints, as well as responding to queries and requests from the research staff and/or collecting data for the experiment. Program operators are concerned that this diversion of staff effort will adversely affect the quality of service the program can provide to its regular clientele.

A closely related concern is that, by diverting some applicants into a control group, random assignment will *reduce*

[3] See Kershaw and Fair (1976) for an example of the level of documentation of the income maintenance experiment treatments.

[4] Even if the experimental benefits were delivered in exactly the same way they would have been in a regular program, it is possible that the experimental staff communicated the incentives embodied in the treatments more clearly--or less clearly--than regular program staff would; this could affect the response to the experimental treatment.

the flow of program participants. For programs that have difficulty recruiting enough participants to fill all their program slots (or as many of a particular subgroup as they would like), this can be a serious concern.

These concerns arise with varying force depending on the experimental setting. They tend to be most strongly felt when the experiment is intended to evaluate an ongoing program. They are less salient in the case of special demonstrations, in part because the experiment is seen as less of a threat to the program's central mission and in part because the funding for special demonstrations normally takes into account the added costs attributable to the experimental design. But researchers are almost certain to encounter these objections, in one form or another, in virtually any experiment that involves regular program staff.

A number of strategies can used by the experimenter to counter these concerns. They fall under the general headings of:

- Establishing the legitimacy and ethical acceptability of the study;
- Minimizing staff burden and cost;
- Ensuring an adequate flow of program participants; and,
- Providing positive inducements for program staff to cooperate.

We discuss each of these general strategies in turn.

Establishing the legitimacy and ethical acceptability of the study.

The researchers' first task in dealing with local program staff is to convince them that the study should be taken seriously — *i.e.*, that it is important and legitimate. In part, this involves convincing them that the researchers themselves should be taken seriously — that they know and understand the program they are evaluating (or within whose context they are proposing to experiment). This means that the members of the research team must do their homework so that they do in fact understand the program. It will be helpful if the research team includes some individuals who have worked in the program or other peers whom program staff respect, such as members of local or national professional associations related to the program.[5] For example, in the National JTPA Study, the site recruiting team included several former directors of local training programs, as well as representatives of the National Association of Counties, the National Alliance of Business, and the National Governors' Association. It is also extremely helpful for the state or federal agency that funds the program to play a prominent role in the initial discussions with local program staff, to underscore the importance they attach to the study.

A necessary part of establishing the legitimacy of an experiment is convincing program staff that it is ethically sound.[6] Doing so requires that staff concerns about denial of service to the control group be taken seriously and addressed directly, through frank and thorough discussion. As discussed in the first paper in this series, there are strong arguments for the ethical acceptability of random assignment in most cases (and, presumably, experiments are only undertaken where such arguments can be made): Most social programs and demonstrations can only serve a fraction of those who are nominally eligible; in such cases, random assignment is arguably the fairest way to ration scarce program benefits or services. Moreover, it can be argued that it is unethical *not* to evaluate ongoing programs with the strongest possible methodology. It is important to know whether the program is achieving its intended objectives; continuing an ineffective program is a disservice to its intended beneficiaries, as well as to the taxpayers who fund it. If, on the other hand, the program is achieving its intended effects — as virtually all program operators believe — the evaluation will provide the information needed to justify its continuation.

Generally, this dialogue must occur twice, at two different levels: with program management staff, who will decide whether the local program is willing to participate, and with the line staff who must implement the experiment.[7] It is important to recognize that, even after program management has agreed to participate, it is essential that the program staff's ethical concerns about random assignment be thoroughly discussed before going on to any other aspect of experimental design or implementation. If they are

[5] This type of program experience and expertise is, of course, valuable in its own right to ensure that the experiment is well-designed, as well as to reassure local program staff.

[6] It may also be necessary and/or useful to demonstrate that random assignment is *legal*. In the National JTPA Study and the National Evaluation of the Food Stamp Employment and Training Program, the sponsoring agencies obtained opinions from the departments' chief legal officers that the agencies had legal authority to employ random assignment to evaluate their programs.

[7] In some cases, several different organizations must agree to cooperate with the experiment. In the National JTPA Study, for example, it necessary not only to obtain the agreement of the local Service Delivery Area, the organization that administers JTPA, but also the consent of the local Private Industry Council, which oversees the program, and, usually, one or more local elected officials, such as the mayor or county commissioners. Any one of these groups could effectively veto the study in the local area. This "multiple veto" problem made site recruiting for the study extremely difficult.

not, these concerns will continue to arise, disrupting the dialogue with the site and making it impossible to focus on implementation of the study. Even if staff are not entirely convinced by the experimenters' arguments, it is important that they have an opportunity to express their views. If given that opportunity, in the end line staff will generally accept whatever decision their superiors have made, even if they do not fully agree with it. The objective here is as much to allow the staff to work through their feelings about denial of service to controls, which can be quite strong and emotional, as to convince them intellectually of the ethical acceptability of the study. In some cases, this takes several meetings with staff; these sessions can become quite heated.

In cases where staff resistance to random assignment threatens to undermine the integrity of the experiment or the site's willingness to participate, it may be necessary to make some modifications to the design to secure site cooperation. For example, the site can be given a limited number of discretionary *exemptions from random assignment*, to be used in cases where, in the staff's view, the applicant is in such dire need of program services that it would be unethical to deny them under any circumstances. The effect of such exemptions, from an evaluation standpoint, is to limit the applicability of the impact estimates to the nonexempt participant population — *i.e.*, the experimental results will not apply to those exempted from random assignment. If the number of exemptions is small, however — say, on the order of one or two percent of all participants — this will have only a negligible effect on the overall impact estimates.[8]

Exemptions from random assignment can also be used to deal with legal barriers to excluding certain individuals from the program. For example, if the program receives court-mandated referrals that it is required to serve, these can be exempted from random assignment. Again, however, it must be recognized that the experimental estimates will not apply to this segment of the participant population.

A second type of design modification that has sometimes been used in evaluations of ongoing programs to overcome staff objections to random assignment is *shortening the "embargo" period during which controls are excluded from the program*. Ideally, controls should be excluded from the program throughout the follow-up period during which outcome data are collected. In some cases, however, realistically there is little risk that controls will reapply after an embargo period of a year or more, so that little is lost by shortening the exclusion period. In the National JTPA Study, for example, controls were excluded from JTPA for only 18 months, even though data collection extended for 30 months after random assignment. Less than one percent of all controls enrolled in JTPA after the embargo period ended; the effect of such a low incidence of "crossovers" on the impact estimates was probably negligible.[9] Caution must be exercised in adopting this type of design modification, however; the threat posed by reapplication of controls before the end of the follow-up period depends on the institutional context, so that each case must be judged on its own merits. Moreover, it must be recognized that if virtually none of the controls reapply after the end of the embargo period, this "concession" is of little value to them.

A third design compromise that has sometimes been used to address staff concerns about the effects of random assignment on controls is to allow intake staff to *provide controls with lists of local providers of services similar to those offered by the experimental program*. This not only assuages staff concerns that needy applicants may go without assistance because of the study; it also allows them to avoid being placed in the uncomfortable position of having to face disappointed controls empty-handed. It may even serve to prevent more active intervention by program staff on behalf of controls, such as individual referrals to specific providers. To the extent that providing lists of alternative providers leads controls to receive services that they would not have obtained in the absence of the experimental program, however, it biases the impact estimates relative to that standard. Because the effects of such information is unknowable in advance, and may be large, we do not recommend this approach. Rather, we recommend that the experiment be designed so that controls are informed

[8] The expected value of the impact estimate in the absence of exclusions is the weighted average of the estimated impact on those excluded from random assignment and the estimated impact on those randomly assigned, where the weights are their relative proportions. Thus, the bias involved in excluding those who are exempted from random assignment is the difference in the impacts on the two groups times the proportion of participants exempted from random assignment.

This bias will be small so long as the proportion exempted from random assignment is small. Suppose, for example, that 2 percent of all participants are exempted from random assignment and that the estimated impact of the program is a 10 percent increase in the outcome of interest. In the extreme case where the program has *no* effect on the outcome for those exempted, the estimated impact will overstate the true impact by 2 percent; i.e., the estimated impact would have been 9.8 percent, rather than 10 percent, if the program's zero impact on the exempted individuals had been included in the estimate.

[9] As with exclusions from random assignment, the bias due to crossovers will be small if their incidence is small. We will discuss the analytic treatment and implications of cross-overs in a later paper in this series.

of their status by letter and program staff have no personal contact with controls after random assignment. This removes both the opportunity and the need for program staff to respond to requests for assistance from controls.

It might be argued that providing no information to controls biases the impact estimates *upward* by artificially reducing the amount of service controls receive below what they would have received in a world without the experimental program. According to this argument, the fact that the controls volunteered for the experimental program is evidence that in the absence of that program they would have sought assistance from some other program. Assignment to the control group may discourage some applicants to the point that they give up on seeking assistance. Thus, the argument goes, a hands-off policy toward controls does not faithfully replicate the desired counterfactual, in which they would all seek assistance.

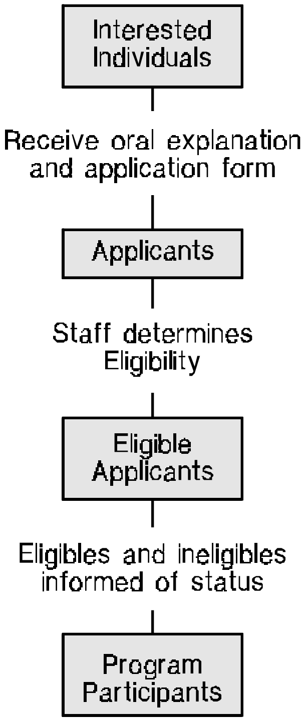
While conceding that this argument has some validity, we would argue that the preferred solution is not to give controls extensive lists of local service providers. Rather, in cases where this argument is persuasive, it is probably better to refer controls to the single source of assistance that they were most likely to contact in the absence of the experimental program. In the case of the JTPA evaluation, for example, it seems most likely that controls would have gone to the Employment Service had JTPA not existed. Thus, referring controls to the Employment Service may provide the conceptually correct counterfactual.

A closely related strategy for assuaging staff concerns about denial of services to controls is to provide some minimum level of service to the control group, rather than attempting to exclude controls from the experimental program altogether. If policy interest focuses on the effects of the experimental treatment relative to no additional service, however, rather than relative to this minimum level of service, this design will underestimate the impact of the experimental program by an unknown amount. Only if the services provided to the controls have no effect — in which case they are really a placebo, not a real treatment — will this design provide unbiased estimates of the incremental impact of the experimental program relative to no additional service. For this reason, if the ethical conditions for a no-service control group are met we do not recommend adopting this approach simply to appease program staff.

Minimizing staff burden and cost.

Normal Program Intake Process

EXHIBIT 1

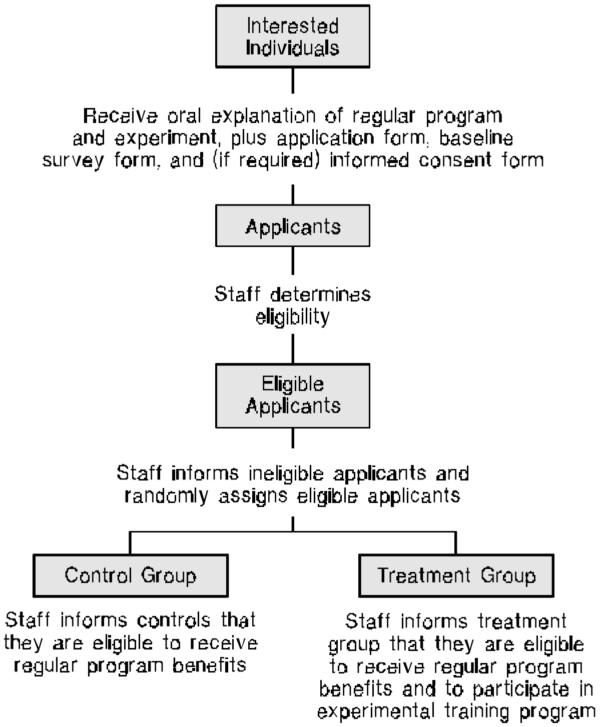

Interested Individuals
Receive oral explanation and application form
Applicants
Staff determines Eligibility
Eligible Applicants
Eligibles and ineligibles informed of status
Program Participants

While less emotional than the issues surrounding denying service to controls, the added staff burden and cost imposed by an experimental study are very legitimate concerns for local program operators. While some added burden and cost are unavoidable, there are several ways in which experimenters can minimize them.

Perhaps the most important way to minimize added burden is to *integrate random assignment and baseline data collection into the regular program intake process with as little disruption of normal program activities as possible*. Generally, implementation of random assignment involves adding four steps to the intake process: (1) informing applicants about random assignment and (if required) obtaining their informed consent; (2) collecting baseline information; (3) randomly assigning the applicants; and (4) informing applicants of the outcome of random assignment. Wherever possible, these steps should be "piggy-backed" onto existing program activities, to avoid unnecessarily complicating and lengthening the intake process.

Consider, for example, an experiment designed to test a new training program for welfare recipients. Exhibit 1 shows the normal intake process for the welfare program: Individuals interested in applying for welfare receive an oral

Program Intake Process with Random Assignment EXHIBIT 2


Interested Individuals
Receive oral explanation of regular program and experiment, plus application form, baseline survey form, and (if required) informed consent form
Applicants
Staff determines eligibility
Eligible Applicants
Staff informs ineligible applicants and randomly assigns eligible applicants
Control Group
Treatment Group
Staff informs controls that they are eligible to receive regular program benefits
Staff informs treatment group that they are eligible to receive regular program benefits and to participate in experimental training program

explanation of the program and its application procedures from program staff. Those still interested in applying for program benefits are given an application form. Applicants complete the form and submit it, along with any required documentation, to program staff.[10] Program staff review the applications and determine eligibility for the program. Those found eligible are notified and informed of any steps they must take in order to receive program benefits; ineligible applicants are informed that they will not receive benefits.

Exhibit 2 shows how random assignment might be integrated into this process. At the initial step, along with an explanation of the regular program, program staff would give individuals who express interest in the program a brief explanation of the experiment and the use of random assignment to select those to be invited (or required) to participate in the experimental training program.[11] The application package given to potential applicants would include a baseline survey form and (if required) an informed consent form, to be completed and submitted along with the regular program application form. Staff would then determine the eligibility of those who apply in the usual manner, and ineligibles would be informed that they do not qualify for program benefits. Eligible applicants would be randomly assigned to treatment or control status. All eligible applicants would then be notified that they qualify for regular program benefits, and those assigned to the treatment group would be invited to participate in the experimental training program (or, if the program is mandatory, informed that they are required to participate).

This integration of random assignment into the intake process has the advantage that it does not change the order or content of any of the regular intake steps. Moreover, it does not increase the number of times that staff must meet with applicants in order to complete the process.

There are other steps that experimenters can take to minimize burden on intake staff. Because intake staff will have to respond to applicants' questions about the baseline survey form, it will be important to *keep the baseline form as short and simple as possible*. Where more complex baseline data are required, it may be necessary for study staff to collect those data through personal interviews with the applicants, rather than through a self-administered baseline form or by asking program staff to interview the applicants. Similarly, there are various options for conducting random assignment, some of which entail less burden on program staff than others. We discuss the options for both baseline data collection and random assignment procedures later in this paper.

In evaluations of ongoing programs, *the random assignment ratio* will be an important determinant of the added cost and burden for program staff. Suppose, for example, that the program normally serves 1,000 participants per year. Random assignment of equal numbers of eligible applicants to the treatment and control groups means that for every 1,000 program participants (treatment group members) the program must process enough applications to produce 2,000 eligible applicants — *i.e.*, the need to provide for a control group doubles the number of applicants that must be processed. If, instead, an assignment ratio of two treatment group members to every control were adopted, only 500 controls would be needed for every 1,000 participants and the required increase in intake would be only 50 percent.

[10] Program staff frequently take down the application information in person, although the applicant is usually required to submit additional documentation, such as birth certificates, proof of residence, etc.

[11] If the participation in the experimental training program is voluntary, it may not be necessary to inform program applicants about it at this point. Those assigned to the treatment group could be informed, and invited to participate, at the same time that they are notified of their eligibility for regular program benefits. If the experimental program is mandatory, prospective applicants must be informed that, if randomly selected, they will be required to participate.

Quite aside from considerations of intake costs, assigning more applicants to the program than to the control group may help to assuage staff concerns about denial of service to controls, because it reduces the number of applicants excluded from the program. In the National JTPA Study, for example, a 2:1 assignment ratio was adopted primarily for this reason.

Such a change in the random assignment ratio is not costless, however. As noted in the previous paper in this series, equally sized treatment and control groups yield the most precise impact estimates (when there are only two experimental groups). Thus, if a different ratio is adopted, the experimenter must either settle for less precise impact estimates or increase the sample size to maintain precision. In the example above, if the assignment ratio is changed from 1:1 to 2:1 and the number assigned to the program held constant at 1,000, so that the number assigned to the control group falls to 500, minimum detectable effects will rise by 22.5 percent. If, instead, the overall sample size (treatment group plus controls) were held constant at 2,000, with 1,333 assigned to the program and 667 assigned to the control group, the increase in minimum detectable effects would be only about six percent.

To achieve the same minimum detectable effects with a 2:1 assignment ratio as with a 1:1 ratio, the overall sample size would have to be increased by 12.5 percent, to 1,500 treatment group members and 750 controls in this example. This increase in sample size would require extending the length of time for which random assignment is conducted by 50 percent, thereby offsetting somewhat the apparent benefit to the program of the higher assignment ratio.[12] More importantly, it would increase data collection costs in proportion to the increase in sample size. For large samples and/or extensive (or expensive) data collection strategies, this can be a substantial cost.[13]

Another important determinant of the added burden posed for intake staff is *the point in the intake process at which random assignment is conducted*. As noted in previous papers, random assignment late in the process means that intake staff have to take applicants who will ultimately become controls through more steps in the intake process. Program staff resist late random assignment for other reasons, as well. Late random assignment makes exclusion of controls more difficult, both for the applicants, whose expectations are raised by prolonged contact with the program, and for the staff, who have become more invested in helping them.

As with changing the random assignment ratio, however, changing the point of random assignment has important analytic costs. Conducting random assignment earlier in the intake process is almost certain to increase the number of "no-shows" — individuals assigned to the treatment group who do not enter the program. While a methodological adjustment is available to remove the effect of no-shows on the impact estimates,[14] that adjustment increases the minimum detectable effects attainable with any given sample size. If the point of random assignment is changed in order to accommodate program staff, then, the experimenter must either accept less precise impact estimates or increase the experimental sample size to maintain the power of the design. This choice can be analyzed in terms very similar to those presented above in connection with a change in the random assignment ratio.

Where the issue is primarily one of the added cost of processing additional applicants, it may be cheaper to reimburse the program for those costs than to increase the sample size sufficiently to maintain the precision of the estimates. In an evaluation of the California Conservation Corps, for example, the program agreed to conduct random assignment after eligibility had been determined only after the study sponsor agreed to reimburse the cost of the physical examinations administered as part of the program's normal intake process, for applicants who were subsequently assigned to the control group.

Ensuring an adequate flow of program participants.

Program staff may also resist random assignment because of the difficulty of recruiting a sufficient number of applicants to fill all program slots *and* to provide for a control group. One response to this problem is to change the random assignment ratio to reduce the number of applicants assigned to the control group. Another is to commit to *temporarily* reducing the random assignment ratio if the program experiences difficulty recruiting enough applicants to fill all its slots. In the National JTPA Study, for example, local programs were required by Department of Labor regulations to spend 40 percent of their training budgets on youth. Many sites had difficulty recruiting enough youth to meet this requirement when one-third of all applicants

[12] When sites in the Minority Female Single Parent Program were given a choice between a 1:1 assignment ratio and a 2:1 ratio with a longer period of random assignment, three of the four sites opted for the 1:1 ratio with a shorter random assignment period (Boruch et al., 1988).

[13] More generally, the cost of adopting a nonoptimal assignment ratio can be measured by the savings (primarily in data collection costs) that would accrue if a smaller, optimally allocated sample with the same minimum detectable effects were adopted.

[14] See the second paper in this series.

were assigned to the control group. To address this problem, the evaluators temporarily changed the assignment ratio from 2:1 to 3:1 or even 6:1 in several sites.

While this change may have kept these sites from dropping out of the study, it posed difficult problems for the analysis, as we will see in the next paper in this series. These same analytical difficulties arise if the random assignment ratio varies across sites. Therefore, if a nonoptimal random assignment ratio is adopted, the experimenter should at least adopt the *same* ratio in all sites.

The experimenters can also provide technical assistance to improve program outreach and/or to reduce the number of applicants who drop out prior to program entry. While the researchers conducting the experiment may not have the expertise to provide such technical assistance, they can usually hire consultants who do.

Providing positive inducements for program staff to cooperate.

In exchange for the added burden and cost that experiments entail, experimenters can offer some positive inducements to participating programs. While many of these benefits to study sites are intangible, they are nevertheless real and should be emphasized in the initial dialogue with prospective sites.

An important inducement for many local program operators is the *opportunity to take part in an important national study*. Local program staff often feel powerless to affect national policy; participation in a study that may influence national policymakers provides them a way to do so. In evaluations of ongoing programs, local program staff may see the study as a way to demonstrate the value of the program. In special demonstrations, they may be attracted by the chance to show the efficacy of new service approaches, in the hopes that they will be funded on a regular basis. As this implies, however, local staff willingness to participate in a demonstration will be strongly conditioned by their view of the desirability of the experimental program being tested.

Local program operators may also be attracted by the *opportunity to obtain information and feedback on their own programs*, even if they do not fully understand or accept the argument for a rigorous experimental evaluation. Most local programs have little systematic information about what happens to their participants after they leave the program. For these program operators, the experiment's follow-up data collection offers a rare opportunity to observe the long-term outcomes of their participants. A commitment to provide such data to the study sites can therefore be an important incentive for local programs to participate.

Experimental studies can also provide an *opportunity for participating programs to learn from each other*. Many multisite experiments hold periodic conferences at which participating sites can share their experiences and discuss issues and problems of common interest. In studies where this type of forum has been provided, local programs have almost uniformly found it very useful.

In many cases, the researchers can also provide valuable *technical assistance to the participating programs*. Much of this assistance is, of course, focused on the implementation of the experiment and collection of data for the study. While this assistance would appear to benefit primarily the experiment, many local programs have found the knowledge of research and evaluation methods gained by participating in an experiment to be quite useful in their own program evaluation efforts. The researchers are also in a unique position to advise local program operators on how other study sites have dealt with difficult operational problems, or at least to put them in touch with knowledgeable staff in other study sites. The researchers may also be able to draw on their own expertise to assist local program staff in dealing with operational problems. For example, expertise in data collection and data processing can often be helpful in setting up program information systems and tracking program participants. Experience acquired in evaluating other programs can often be applied to problems such as recruiting program applicants or reducing attrition among applicants or participants.

Unfortunately, the value of some of these relatively intangible benefits may only become apparent once the study is underway. A more obvious, and sometimes more convincing, inducement to participate is *monetary reimbursements*. In the National JTPA Study, for example, the Department of Labor made payments averaging $170,000 to the local Service Delivery Areas to compensate them for study-related costs.[15] While it is appropriate to reimburse local programs for the added costs associated with the experiment, researchers should take care that such payments do not devolve into simple bribes, lest prospective study sites adopt a strategy of withholding their agreement to participate in order to maximize the monetary payment.

Where the experiment is being conducted by the federal or state agency that administers the program, the agency can also provide some nonfinancial inducements to participate. For example, in programs with performance standards systems, study sites might be held harmless with respect to any adverse effects that participation in the experiment might have on their performance indicators. Study

[15] See Doolittle and Traeger (1990).

sites might also be given more flexibility in meeting other program regulations during the study period, in recognition of the special demands of the experiment.

Implementing Random Assignment

The actual assignment of potential participants to treatment and control groups is quite straightforward. In virtually all modern experiments, this is done by a computer algorithm that assigns each potential participant to an experimental group on the basis of a random number generated by the computer.[16] The more complex and difficult aspect of implementing random assignment in a real-world program is establishing procedures to be followed by the research and program staffs that allow the assignment to be made without delaying the program intake process or unduly inconveniencing applicants. Moreover, random assignment procedures must be designed to ensure that intake workers cannot manipulate the assignments to admit favored applicants to the program, and that program staff do not inadvertently subject sample members to the wrong treatment. Finally, the process must be designed to accurately capture certain information that will be critical to later data collection and analysis — principally, permanent identifiers for each individual (*e.g.*, name and Social Security number), the date of random assignment, and the assignment itself.

Two principal approaches have been developed to achieve these objectives: centralized random assignment and on-site random assignment. We discuss each in turn. In the final part of this section, we discuss the random assignment algorithm itself.

Centralized random assignment.

In all of the early social experiments, random assignment was conducted centrally by the research staff in charge of the study. In studies like the income maintenance experiments where the sample was identified through household surveys in the study sites, sample members were randomly assigned as their completed baseline interviews were returned to the study office. Research staff then recontacted those families and invited them to participate in the experiment.

This approach had the advantage of maximizing the experimenters' control over the random assignment process. It provided tamper-proof assignments that were accurately implemented, since all the steps in the process were under the control of research staff. Reliable identifying information for each sample member was obtained from the baseline interview, and the assignment and date of assignment were recorded in the study computer at the time the assignment was made.[17]

Because it relied on obtaining hard copy interviews from the field, this assignment process could take days or weeks to complete. But such delays were not an important consideration in these experiments, because the potential participants had not applied for any program and were not expecting any further contact beyond the interview.

In later experiments, where the experimental sample was drawn from applicants to special demonstrations or ongoing programs, the time required to conduct random assignment became a more critical issue. Program staff strongly resisted putting applicants "on hold" for days or weeks while the researchers conducted random assignment. In response to this concern, a number of later experiments adopted a procedure in which random assignment was conducted over the telephone. When local program staff had completed their eligibility determinations for one or more applicants, they would call a specially designated random assignment clerk and submit their identifying information. The clerk would enter this information into the computer, obtain the assignment, and inform the intake worker of the individual's experimental assignment.

This process was substantially faster than communicating with the site by mail. But it was also very error-prone; oral transmission of names and Social Security numbers led to numerous errors in these critical identifiers. In nearly all cases, these errors could be corrected later by comparison with the hard copy baseline interview forms, but the reconciliation process was a time-consuming, expensive one. Moreover, to maximize responsiveness, the researchers had to have full-time random assignment clerks standing by to take calls; for smaller studies, this was inordinately expensive. A compromise arrangement, which was somewhat less responsive but much cheaper, involved allotting each study site specific hours within which to call in for assignments.

[16] There are, however, a few exceptions to this rule. In one case, the first assignments were made by literally flipping a coin because the computer algorithm was not yet ready for use. In another study, assignment was made by drawing names from a hat, in the presence of the potential participants, in order to demonstrate the fairness of the process.

[17] This approach had the further advantage that it ensured that baseline data were available for all sample members, since they could not enter the study sample (i.e., be randomly assigned) until the baseline interview had been received. It also ensured that the baseline data were not collected *after* random assignment; as we will see later in this paper, this is an important consideration in the experimental analysis.

To avoid the errors induced by oral transmission of information, in some recent studies program staff have transmitted sample identifiers and received the resulting experimental assignments by fax, rather than telephone. This approach is particularly suitable when applicants are recruited and assigned in batches. For example, many training programs recruit participants in "waves", in order to fill a specified number of slots in a training class starting on a specific day. In contrast, welfare programs accept and process applications continuously. Both telephone and fax transmission are somewhat cumbersome when applicants are to be randomly assigned individually.

On-site random assignment.

In cases where quick turnaround is important, the fastest, most flexible approach is to allow program staff to conduct random assignment on site. Until recently, however, on-site random assignment was quite error-prone and potentially subject to manipulation by program staff.

In the earliest studies to use this approach, random assignment was based on the applicant's Social Security number. For example, all applicants with Social Security numbers ending in an even digit might be assigned to the treatment group and those with numbers ending in an odd digit assigned to the control group.[18] Faithfully implemented, this algorithm would indeed generate two well-matched groups, since the last four digits of the Social Security number are, for all practical purposes, randomly assigned to individuals.

Such an algorithm is easily manipulated by local program staff, however, by the simple expedient of misreporting Social Security numbers. By changing the last digit of the reported Social Security number from odd to even, program staff can ensure that favored applicants — *e.g.*, those deemed most needy or simply most likely to complain if assigned to the control group — are admitted to the experimental program. Even if they do not deliberately falsify information in this way, the fact that each applicant's treatment status can easily be known from the beginning of the intake process may lead intake staff to treat those destined to be controls differently from those who will be allowed to enter the program. For example, they might discourage (or simply fail to encourage) those with odd numbers from filling out long application forms or they might be less diligent about determining the eligibility of those whom they know will be controls.

Some protection from these threats to random assignment can be obtained by making the random assignment algorithm somewhat more complex and giving responsibility for its application to a single "random assignment coordinator" in each site — in effect, conducting centralized random assignment within the site. For example, each two-digit number between 00 and 99 could be randomly assigned to treatment or control status and a list of these assignments given to the random assignment coordinator. Individual intake workers would then submit applicants' Social Security numbers to the random assignment coordinator, who would use the last two digits of the number to obtain their assignments from this list. As long as the random assignment coordinator can be trusted not to photocopy the list and hand it out to the intake workers, this approach provides reasonable protection against manipulation of the assignment and/or asymmetric treatment of the different experimental groups — at the cost of some loss of flexibility and rapid turnaround.

An alternative method of on-site random assignment that provides protection against staff manipulation can be employed in some experiments. When the existing participants in a program are to be randomly assigned, the assignments can be conducted, and the experimental status of each individual recorded, in the program's central computer system. This approach could be used, for example, when a new service or requirement for the existing AFDC caseload is to be tested. Randomly assigning the entire caseload*en masse*, using a computerized algorithm, rules out manipulation by program staff to assign specific individuals to the treatment group.

This approach may have other drawbacks, however. To maintain well-matched treatment and control groups, all individuals who were randomly assigned must be included in the experimental analysis. If the experimental treatment is voluntary or if only a subset of participants are eligible to receive it, randomly assigning all participants rather than only those who volunteer and/or are eligible will result in a higher proportion of nonparticipants in the treatment group. Thus, the estimates of impact on participants will be less precise.

The advent of virtually universal availability of microcomputers has opened the way for more flexible, yet secure, methods of on-site random assignment. Software that will run on virtually any microcomputer is now available to allow individual intake workers to randomly assign individual program applicants in much the same way as the random assignment clerk in a centralized random assignment proce-

[18] Other random assignment ratios are readily implemented by using the last two digits of the Social Security number. For example, a 2:1 ratio can be achieved by assigning all those whose Social Security number ends in "67" or less to one group and those whose last two digits are "68" or greater to the other.

Blocked Random Assignment: Illustrative Assignments (Block Size = 6)

EXHIBIT 3

APPLICANT NUMBER	1	2	3	4	5	6	7	8	9	10	11	12	13	
ASSIGNMENT	T	T	C	T	C	C	C	T	T	C	C	T	T	

<——————— BLOCK 1 ———————> <——————— BLOCK 2 ———————> ...

dure.[19] The intake worker need only enter the identifiers needed for data collection and analysis purposes (usually the applicant's name, Social Security number, and date of birth). The software will check to see if this person has been assigned previously and immediately return the individual's assignment on the screen.[20] The software automatically records the identifiers, assignment, and date in an encrypted file that cannot be changed by site staff. This file is periodically submitted to the researchers, either electronically or on a diskette, to create the initial record for each person randomly assigned. On command, the software will also print lists of treatment and control group members for use by program staff. This approach allows random assignment to be quickly and easily conducted at the agreed upon point in the intake process, with virtually no risk of manipulation or tampering by program staff and minimal risk of error in the sample identifiers or recorded treatment status.

It must be recognized that any method of random assignment — centralized or on site — is subject to the risk that program staff will erroneously assign the wrong individuals or assign them at the wrong point in the intake process. We return to these risks later in this paper.

The random assignment algorithm.

As noted above, in modern experiments random assignment is generally based on a computerized algorithm that uses random numbers to assign individuals to groups. The simplest approach would be to draw a random number for each person to be assigned, with the assignment depending on the range within which the number falls. For example, to assign treatment and control group members in a 1:1 ratio, one could draw a random number between 0.0 and 1.0 and assign all those with numbers greater than 0.5 to the treatment group and all those with numbers 0.5 or less to the control group. This is the computerized equivalent of flipping a coin for each person to be assigned.

This approach will produce two well-matched experimental groups, but it has a significant disadvantage. Just as flipping a coin repeatedly will sometimes generate long strings of consecutive heads or long strings of consecutive tails, randomly assigning each individual independently will sometimes produce — purely by chance — long strings of consecutive assignments to the same group. While this is not a problem for the analysis, it can sometimes create operational problems. For example, program staff may lose faith in the fairness of random assignment if they see a large number of applicants assigned to the control group — especially if it occurs early in the experiment.

Assignment of unbalanced numbers of individuals to the two groups may also complicate program planning. Suppose, for example, that the experimental treatment is a training program and that the program staff wants to start a class of ten trainees. With a 1:1 assignment ratio, they would expect to just fill the class if they submitted 20 names for random assignment. But with independent random assignment of each person, they would sometimes receive 15 assignments to the treatment group and sometimes receive only 5.

Relatively close balance in the numbers assigned to each experimental group can be assured by a technique known as **blocked random assignment**. In blocked random assignment, short "blocks" of assignments are created in advance. Within each block, the numbers of individuals assigned to the various groups exactly equal the random assignment ratio, but the order in which they occur is random. Exhibit 3 shows a schematic diagram of a set of such

[19] The software described here was developed by Abt Associates for use in the evaluation of the Head Start Family Service Centers, the National Community Service evaluation, and the Moving to Opportunity Demonstration.

[20] For individuals who have been previously assigned, the software simply returns the same assignment.

blocks. In this example, each block contains six assignments, divided equally between the treatment and control groups.[21] These blocks of assignments are then stored in the computer, in order, and each person submitted for random assignment receives the next assignment in the block.[22]

Under this procedure, at the end of each block the ratio of assignments to the two groups exactly equals the intended random assignment ratio. Within each block, the relative number of assignments to the two groups cannot differ from the desired ratio by more than three persons (if the block size is 6).[23] And since the assignments were randomly ordered within blocks, assignment to the two groups is entirely random.

Blocked random assignment does raise the theoretical danger of intake staff being able to manipulate random assignment. Program staff cannot, of course, see the upcoming assignments that are encoded within the computer. If, however, an intake worker knew that the assignments were in blocks of 6, he or she would know that anytime the numbers of treatment and control group members differed by 3, the next assignment would be to the group with fewer members. Thus, by holding a favored applicant until there were three more controls than treatment group members, he or she could ensure that applicant of admission to the program. In our experience, program staff are neither this analytic nor this determined to subvert random assignment. In any case, if this type of strategic behavior is considered to be a realistic threat, it can be prevented by simply using blocks of different length, in random order.

In implementing a system of blocked random assignment, it is important to think about the organizational level at which balanced assignments are desired. Blocked random assignment assures only that the assignments made by a single computer are balanced. If those assignments are distributed across multiple sites or offices, there is no guarantee of balance within an individual site or office. Alternatively, if multiple computers are used to make assignments within a single site or office, there is no guarantee of balance at the overall site or office level. Generally, operational considerations argue for balanced assignments within each program office. This means that the optimal arrangement will generally be to use one computer in each office. If random assignment is centralized, separate assignment sequences can be reserved for each office.

Maintaining the Integrity of the Experimental Design

Careful design of random assignment procedures will go a long way toward ensuring that the experimental design is implemented as intended. However, as with any field undertaking, a number of things can go wrong in the implementation of an experiment and the researcher must anticipate the potential threats to the design, so that implementation problems can be avoided or, at least, detected and dealt with promptly. The major potential threats to the design that arise in the field include:

- Nonrandom assignments, as a result of deliberate or inadvertent subversion of random assignment procedures by program staff or problems with the random assignment algorithm;
- Random assignment of ineligible program applicants;
- Failure to maintain adequate records of all individuals randomly assigned;
- Controls receiving an amount or kind of service or benefit that they would not have received in the absence of the experiment (“control group contamination”); and,
- Failure to serve treatment group members.

All of these threats involve some degree of lack of understanding of, or cooperation with, the experimental procedures by the local program staff. One general prescription for avoiding these kinds of problems, therefore, is a concerted up-front effort to gain the willing cooperation of the staff and to train them thoroughly in the experimental procedures. As the experiment proceeds, the researchers must also monitor events in the field closely, to ensure that the prescribed procedures are being followed, and provide any necessary retraining. Provision of a written procedures manual for staff reference and availability of research staff to respond to program staff questions will also help ensure adherence to correct procedures.

Nonrandom assignment.

If the computerized random assignment techniques described in the previous section are employed — on either a centralized or decentralized basis — there is little chance that program staff can manipulate, or inadvertently dis-

[21] If the random assignment ratio were 2:1, rather than 1:1 as in the exhibit, each block would contain 4 treatment group assignment and 2 control assignments, in random order.

[22] Equivalently, the blocks can be created by the computer as needed. The essential point is that assignments are created in blocks containing a fixed ratio of assignments to the different groups, rather than one at a time.

[23] This would occur if the first 3 assignments in the block were all to the same group.

tort, the actual random assignment process. In those circumstances, the only real threat of nonrandom assignments is a malfunction of the random assignment algorithm itself. Given the critical importance of random assignment to the success of the experiment, however, it is prudent to guard against even this unlikely event.

In experiments that rely on noncomputerized random assignment — especially where it is conducted on-site by program staff — the risk of nonrandom assignment is very real. For example, in a pilot test of an evaluation of the Women, Infants, and Children's feeding program (WIC), local staff recruited women in health clinic waiting rooms and randomly assigned them on the spot, using an algorithm based on the woman's Social Security number. Proper application of the algorithm would have produced equally sized treatment and control groups. In one site, nearly two-thirds of the women were assigned to the treatment group; it seems clear that recruiters falsified Social Security numbers to allow women who should have been controls to be assigned to the program.[24]

Several steps can and should be taken to detect departures from random assignment. First, the numbers of treatment and control group members assigned should be closely monitored. As explained in the previous section, under blocked random assignment, the numbers in the two groups should never differ from the desired random assignment ratio by more than a specified number. For example, if the random assignment ratio is 1:1 and the block size is 6, the numbers in the two groups should never differ by more than three (one-half the block size). Any larger difference would be a clear indication of a breakdown in the random assignment algorithm. In experiments where blocked random assignment is not used, the indications are not so clear-cut — virtually any difference in group sizes *could* be produced by pure chance. But large deviations from the intended ratio are unlikely and should be treated as strong indications of a failure of random assignment.

Second, as baseline data on the sample become available, it is possible to compare the characteristics of the treatment and control groups, to determine whether they are well-matched. Again, any degree of mismatch between the two groups is possible due to chance alone. But a finding of more statistically significant differences between the two groups than would have been predicted on the basis of chance alone (*e.g.*, significant differences at the 10 percent level on more than one out of ten characteristics) should be treated as an indicator of potential problems. If either of these checks sound a warning signal, the experimenters should review the entire random assignment process, including both site procedures and the random assignment algorithm, to determine whether there has been a breakdown. In any case, the researchers should periodically conduct on-site reviews of local program procedures, in order to ensure compliance with the experimental design.

Random assignment of ineligible individuals.

One of the most frequent departures from the experimental design, especially early in the experiment, is random assignment of ineligible individuals.[25] This can happen for a variety of reasons. For example, staff may randomly assign individuals before an eligibility determination is made, either inadvertently or because they misunderstand the correct timing of random assignment. Even when the individual has been determined to be eligible prior to random assignment, sometimes information subsequently comes to light that proves that determination to have been in error. Program staff may also randomly assign individuals previously determined to be ineligible — again, either inadvertently or because they misunderstand the random assignment procedures.

Ineligibles who have been randomly assigned by mistake are usually detected — if they are detected at all — by program staff in the course of their normal program activities. For example, a case worker may make the discovery in the course of an initial counseling session with a client to determine their service needs. This discovery is generally followed by a frantic call to the research staff, asking, "What do we do now?"

Several approaches are available to deal with random assignment of ineligible individuals. If the ineligibles assigned to *both the treatment and control groups* can be identified, they can be removed from the experimental sample — in effect, "unassigned." Removal of all ineligibles from both groups leaves two well-matched groups of eligibles, just as if the ineligibles had never been assigned in the first place.

It is critical, of course, that the ineligibles in both groups be identified if this approach is to be taken. Frequently, the process that leads to detection of ineligibles applies

[24] See Puma et al. (1991).

[25] In this context, by "ineligible individuals" we mean individuals who are *ineligible for random assignment*. If random assignment is conducted after eligibility for the program is determined, these individuals are also ineligible for the experimental program. In experiments where random assignment is conducted early in the intake process, however, individuals who are ineligible for the program could be eligible for random assignment.

only to the treatment group. For example, in one site of the Moving to Opportunity demonstration, counselors working with the experimental group assigned to receive housing vouchers that could only be used in low-poverty areas found that 20 of the first 148 families assigned to that group (14 percent) were ineligible. Since the experimental groups should be well matched on all characteristics, one can infer that a similar proportion of those assigned to the other experimental groups was ineligible. However, among those assigned to unrestricted housing vouchers, who received no counseling, program staff identified only seven percent as ineligible, and they found no ineligibles among those assigned to the control group, with whom program staff had no further contact. Clearly, the probability of detection of ineligibles varied with the amount of staff contact with the families.

As this example suggests, one can usually get a good idea whether the ineligibles in all experimental groups have been identified by examining the relative numbers of known ineligibles in each group and the process by which they were discovered. If the proportion of sample members known to be ineligible differs substantially among groups and/or the process by which they were detected seems more likely to identify ineligibles in one group than in another, they should not be removed from the sample.

Even when ineligibles cannot be removed from the sample, an analytic correction is available if the ineligibles in the treatment group can be identified and excluded from the program before they receive any services. In that case, they can usually be treated as no-shows, and the no-show correction described in a previous paper in this series can be applied to remove their effect on the impact estimates. The only assumption required for this correction to yield unbiased estimates of program effects on eligible participants is that the program have no effect on the outcomes of the no-shows in the sample, including the ineligibles. Thus, even in those instances where it is unlikely that ineligibles can be identified in the control group, it is important to attempt to identify any ineligible individuals in the treatment group as early as possible.

If the assumption required to apply the no-show correction is not satisfied, the experimenter has no choice but to include the ineligibles in the sample and to recognize that the resulting impact estimates apply to a somewhat different population than the intended eligible population.

Failure to maintain adequate records.

To conduct a valid experimental analysis, we must know the identity of all individuals assigned, when they were assigned, and the group to which they were assigned. If random assignment is conducted using a computerized algorithm designed by the researchers, as suggested in the previous section of this paper, a record containing these pieces of information is automatically created at the time of assignment.

When the sample is assigned by noncomputerized methods, however, errors and omissions can creep into these data, especially when local program staff are responsible for random assignment. Program staff have a natural tendency to focus on program participants; they may keep only minimal records of those assigned to the control group or of treatment group members who do not enter the program. Moreover, they may be very careless about retaining the records of these latter groups; unless the researchers collect this information soon after random assignment, it may be lost.

In an experiment to test the delivery of family support services by local nonprofit agencies, for example, random assignment was conducted by the local programs before the evaluation contractor was selected. Later, when the evaluation contractor attempted to collect random assignment information, some of the sites could not produce accurate lists of assignments.[26] In these cases, evaluation staff had to search through hard copy records in the local program office to reconstruct the assignments. In one site, the researchers were simply unable to identify the complete sample of families who were randomly assigned; that site had to be dropped from the experiment.

Similar experiences in other experiments where the design and implementation of random assignment were entrusted to local program staff with little or no oversight by the research staff lead us to the conclusion that random assignment should always be designed and supervised by researchers. While local staff may actually perform the assignments — as in the decentralized random assignment approaches described in the previous section of this paper — they should do so only after thorough training, using systems and procedures designed by the research staff.

Control group contamination.

Perhaps the most difficult problem to deal with after the fact is controls receiving an amount or type of service that they would not have received in the absence of the experi-

[26] In this experiment, bad recordkeeping was compounded by a complex design. At the sponsoring agency's suggestion, most of the sites had originally assigned families to three groups: a treatment group, a control group, and a "replacement group", which was intended to be a source of families to replace those who failed to enter the program or dropped out after entry. In some cases, families had been nonrandomly selected from this latter group to replace no-shows and dropouts; these families had to be identified and excluded from the analysis.

mental program. It is important to recognize that the mere fact that controls receive some nonexperimental services does not necessarily mean that control group contamination has occurred. In cases where services similar to the experimental treatment are available from nonexperimental sources, the desired counterfactual involves *some* receipt of services by the control group.[27] Because we do not know the *amount or type* of services the control group would have received in the absence of the experiment (that, after all, is the purpose of the control group!), it is virtually impossible to measure control group contamination once it has occurred, and therefore virtually impossible to correct for any resulting bias. This means that it is critically important to prevent control group contamination from occurring in the first place.

Control group contamination can arise from several sources. Program staff may refer controls to other sources of assistance, either out of a simple desire to be helpful or as a "consolation prize" for having been excluded from the experimental program. The mere fact of having been recruited for, then excluded from, the experimental program may change controls' behavior in seeking similar services from nonexperimental sources. On the one hand, outreach for the experimental program may prompt some individuals who would not have sought services at all to seek them. Once assigned to the control group, some of these individuals may go on to seek nonexperimental services. In these cases, the existence of the experiment encourages a higher rate of service receipt than controls would have experienced in its absence. On the other hand, assignment to the control group may discourage some individuals from seeking nonexperimental services, even though they would have in the absence of the experiment. Thus, we cannot even be certain of the direction of any bias.

Little can be done to prevent any contamination that occurs simply as a result of program outreach. A number of steps can be taken, however, to prevent program staff from taking actions that may contaminate the control group. Perhaps the most effective way to protect against program staff undermining the experimental design by helping controls obtain nonexperimental services is to make sure that the staff understand and accept the reason for this prohibition. In our experience, the greatest threat of control group contamination comes from staff who either do not understand what they can and cannot do, and therefore violate the prohibition unintentionally, or feel that random assignment is unfair and deliberately attempt to compensate for what they see as an injustice to controls. Therefore, in training local program staff in random assignment procedures, it is important to be clear about the prohibition on special efforts to help controls and to take the time to work through staff concerns about random assignment to the point where they feel comfortable abiding by this prohibition.

An effective way to remove both the opportunity and the temptation for program staff to refer controls to nonexperimental services is to inform controls of their experimental status *by letter*, rather than in person. This has the added advantage of ensuring that each control receives a full, clear explanation of the reason for exclusion from the program and their rights to receive services from other sources (without identifying those sources). It also creates a permanent record of what the controls were told.

The opportunity for control group contamination can also be reduced by minimizing contact between program applicants and service providers *prior* to random assignment, especially where experimental services are purchased from outside vendors. Service vendors often have funding from multiple programs; in these cases, well-meaning staff may tell applicants that if they are assigned to the control group the provider will serve them under another program. It is therefore preferable for the central administrative organization for the experiment to conduct intake and random assignment, rather than entrusting those functions to service providers.

It is, of course, also necessary to take steps to prevent controls from receiving the experimental treatment. Here again, the willing cooperation of program staff, and their thorough understanding of the prohibition on services to controls, is essential. Given those, perhaps the greatest risk of controls receiving experimental services arises from controls who reapply to the experimental program. The random assignment software can be programmed to detect those who reapply while random assignment is still ongoing.[28] If the program continues to accept applicants after the end of the random assignment period, the embargo on services to controls can be enforced by creating a record for each control in the program's central participant infor-

[27] See the discussion of "incremental impacts" in the second paper in this series.

[28] If random assignment is conducted on a decentralized basis, however, this will detect only those who reapply to the same office. In these cases, separate procedures must be established to check each applicant against a master list of controls, as suggested below for those who reapply after the end of the random assignment period.

One way to eliminate the need to check for previous assignments while still ensuring that individuals receive the same assignment when they reapply is to use a random assignment algorithm based on the applicant's Social Security number. Such an algorithm will always return the *same* assignment to those who reapply. Unfortunately, assignment on the basis of the applicant's Social Security number is incompatible with "blocked" random assignment, discussed earlier in this paper.

mation system, with a flag or note indicating that this person should not be accepted as a participant until after the end of the embargo period.

In a subsequent paper, we will discuss an analytic correction that can be used to correct for any "crossovers" (controls who receive the experimental treatment) that do occur. Here, we note only that crossovers should *not* be dropped from the experimental sample, as that will unbalance the treatment and control groups. Rather, they should simply be identified in the data base and included in all regular data collection activities, so that the analytic correction can be applied.

Failure to serve treatment group members.

Just as it is important that controls not receive services that they would not have received in the absence of the experiment, it is important that the treatment group receive the experimental treatment. As noted earlier in this paper, the presence of individuals in the treatment group who do not receive the treatment degrades the precision of the estimates. A 30 percent nontreatment rate has the same effect on minimum detectable effects as a 50 percent reduction in sample size; a 50 percent nontreatment rate is equivalent to losing three-fourths of the experimental sample.

There will, of course, be some no-shows in virtually any program. Some individuals will change their minds after being accepted into the program, or their situation will change so that they no longer need, or cannot take advantage of, program services. But high rates of nontreatment are probably a sign of problems in the administration of the program or of random assignment itself.

Perhaps the most common reason for nontreatment of those randomly assigned to enter the program is long lags between random assignment and entry into the program. This typically occurs because program applicants are assigned before the program is ready for them. For example, early in the AFDC Homemaker–Home Health Aide Demonstration, several sites started recruiting potential trainees well in advance of the scheduled start of training and randomly assigned applicants as they were approved. This resulted in long lags between assignment and the start of training for the earliest applicants, and many of them dropped out before entering training. In a few cases, inadequate numbers of applications caused classes to be postponed, again resulting in long lags and high no-show rates. These problems were resolved by better coordination of recruiting activities with the start of training and by adopting a practice of holding applications until just before the class was about to begin, then reconfirming the applicants' availability before randomly assigning them.

As this example suggests, high rates of nontreatment can often be reduced by improved management of the intake flow or by changing the timing of random assignment to reduce the lag between random assignment and program entry.

In some cases, high rates of nontreatment arise because the random assignment model itself is flawed or severely constrained by institutional factors. If, for instance, all program applicants are randomly assigned rather than only those who are eligible for the program, the resulting treatment group will contain a number of individuals who do not participate in the program because they are ineligible. Sometimes this is unavoidable. For example, some programs employ extensive testing or in-person screening as part of the eligibility determination process; in these programs, it would be extremely expensive and/or burdensome (for both staff and applicants) to carry all applicants through this process before randomly assigning a large fraction of them to the control group.[29] While such circumstances will usually dictate placing random assignment before eligibility determination, the experimenter should be cognizant of the costs of doing so, in terms of reduced precision of the estimates and/or the need for a larger sample to maintain precision.

Data Collection

Several different types of data will be required in the experimental analysis. Perhaps the most fundamental data element is the **treatment status indicator**, which shows the experimental group (treatment or control) to which each sample member was assigned. This information is needed to create the experimental contrast that measures program impacts. The **date of random assignment** is required in order to align outcome data in time across the sample. **Personal identifiers** will be needed to link data from various sources together. **Baseline data** on sample members' background and demographic characteristics are useful for describing the study population, improving the precision of the impact estimates, and defining subgroups of the study sample for separate analysis. Data on the **outcomes** of interest are needed in order to estimate the impacts of the experimental program. Data on **program participation** (including nonexperimental services received by

[29] Applicants to some youth corps, for example, are required to complete a try-out period that can last as long as four weeks, before they are accepted as "appropriate" for the corps. Up to a quarter of the applicants who begin this try-out period drop out before completing it.

both the treatment and control group) are required to determine the service differential that produced the estimated impacts. Finally, measures of the **cost** of services received by the treatment and control groups are necessary in order to allow comparison of program impacts with the net additional cost of the experimental services.

In this section, we discuss the collection of each of these types of data. We consider first the information generated by the random assignment process, then discuss the collection of baseline data, outcome data, program participation data, and cost data. The discussion here is confined to those aspects of data collection that are directly related to the experimental nature of the study; we do not attempt to provide a comprehensive guide to the collection of these various types of data.

Two general rules apply to all data collection activities in an experiment. First, to the extent possible, all data elements should be collected for every individual randomly assigned. Experimental impact analysis involves comparison of the outcomes of the *entire* treatment group with those of the *entire* control group. If any sample members are eliminated from the sample, these two groups will no longer be well matched and, therefore, the impact estimates will be biased. Thus, for example, treatment group members who fail to participate in the experimental program, or controls who do participate, should *not* be dropped from the sample.

Second, as a general rule, the *same* data collection procedures should be used for the treatment group as for the control group. If different methods are used for the two groups, differential errors in the data may be confounded with the impact of the experimental treatment.

We elaborate on the application of these two rules to specific types of data collection in the following sections.

The Random Assignment Record

As mentioned earlier, the random assignment process must be designed to record certain key pieces of information about each person randomly assigned: the treatment status indicator, the date of assignment, and at least one personal identifier. We refer to these data collectively as the **random assignment record**.

These data elements are perhaps the most critical pieces of information about the sample that will be collected or generated in the experiment. All of the impact estimates rest squarely on knowledge of the group to which individuals were assigned, and all of the follow-up data will be keyed to the date of random assignment. Personal identifiers are the key to linking data from different sources, including linking treatment status to outcome data; without this link, the other data collected by the experiment are useless. Sample members missing any of these three critical data elements usually have to be dropped from the analysis.

The use of personal identifiers is illustrated by a typical random assignment and data collection process. One or more personal identifiers (*e.g.*, name and/or Social Security number) are entered into the random assignment record at the time of assignment. Those same identifiers are entered on a separate baseline survey form, which contains background information on the sample member and contact information (*e.g.*, address and telephone number) that can be used to conduct follow-up interviews. The same identifier may be used to access data from the experimental program's records for the person, as well as other administrative records (*e.g.*, AFDC benefits). Each sample member's personal identifier(s) is then used to link all of these data into a single record, so that their background characteristics and treatment status can be related to their outcomes in estimating program impacts.

Fortunately, when random assignment is computerized, the treatment status indicator and date of random assignment are generated by the computer itself. So long as the assignment algorithm is working properly and the computer's internal clock is set correctly, these data will be error-free.

Personal identifiers can be more problematic. While it may seem a simple matter to obtain a sample member's name or Social Security number, these basic data can be remarkably error-prone. Names change (as when women marry) and individuals may use different names at different times (*e.g.*, nicknames or informal versions of their given name). Names may also be misspelled or incorrectly transcribed by program staff, especially if the sample member's handwriting is illegible. Moreover, more than one sample member may have the same name. For all these reasons, names do not serve well as primary identifiers, especially in large samples. They are, however, useful as a secondary identifier to resolve ambiguities that arise is using other identifiers. (Date of birth is also useful for this purpose.)

The ideal identifier would be permanent and unique to the individual. The identifier in common use that comes closest to these characteristics is the Social Security number. The Social Security number has the added advantage that it is used as the primary identifier in many administrative record systems, such as AFDC and food stamp benefit records and Unemployment Insurance wage records. Thus, if the experiment needs to access outcome data from these systems, the Social Security number is the identifier of choice. Although in practice there can also be problems with this identifier (*e.g.*, individuals with multiple Social Security numbers or none at all, transcription errors, etc.), it is widely used as an identifier in many research projects.

Baseline Data

As noted above, baseline data on the sample serve several purposes. First, they are needed to *describe the study population*. In future policy applications of the experimental results, it will be important to know how closely the experimental sample resembled the population for whom the program is being considered. Second, baseline data can be used to *improve the precision of the impact estimates*. (We will discuss the procedure for doing so in a subsequent paper.) Third, baseline data can be used to *define subgroups of the experimental sample for separate analysis*. For example, it is often of interest to know whether program impacts differed between men and women, older and younger participants, those with high school diplomas and those without, etc. This information can be used to target the program on those for whom it is most effective and/or to identify those populations for which the program needs improvement. Finally, if follow-up surveys are to be conducted to collect outcome data, *contact information* must be collected at baseline. Contact information includes the address and telephone number of the sample member, as well as those of friends or relatives who will know how to contact the sample member if he or she should move or change telephone numbers.

The kinds of baseline data required in any particular experiment can be derived from these analytic uses. They obviously include any personal characteristics or experience that would be useful in describing the sample for policy purposes or in defining subgroups of interest for policy. For example, in an experimental test of a job training program, one would want to collect data on the applicants' work experience and education level, as well as standard demographic information such as age, gender, and ethnicity. For purposes of improving the precision of the estimates, the most useful baseline variables are preprogram values of the outcomes of interest (*e.g.*, employment and earnings) and any personal characteristics that can be expected to affect those outcomes.

In designing baseline data collection instruments, it is important to bear in mind the relatively limited role these data play in experimental analyses. A common error is to collect extensive retrospective data of the type one might use in a longitudinal study to "model" the development of the outcome variables over time. In an experiment, the control group provides a longitudinal picture of how the outcomes would develop in the absence of the experimental program, and the treatment–control difference in outcomes tells the researcher how the program changes that picture. While there may be independent interest in how various background characteristics of the individual affect the outcomes, it is not necessary to estimate those effects in order to measure the impact of the program.

Generally, the best source of baseline data is the applicants themselves. Information can be collected directly from the applicants as part of the intake process, in several different ways. Perhaps the simplest, most efficient way to collect baseline information is through the use of a self-administered form, to be completed by the applicant along with the regular program application form. In some cases, however, where the information required is too complex and/or the applicants' literacy level too low to allow use of a self-administered form, it may be necessary to collect baseline data through a personal interview. This can be done either by program intake workers or by interviewers employed by the experimenters. Generally, when personal interviews are required, it is highly preferable that they be conducted by interviewers employed and trained by the researchers, unless the program staff to whom this responsibility is assigned are thoroughly trained by the research staff and are fully committed to the study.

A middle ground between the self-administered form and the personal interview is the "staff-assisted" baseline form, which applicants are asked to complete in the presence of program staff, who are available to respond to questions and explain any parts of the form that the applicants have trouble with. Because staff-assisted forms can be administered to groups of applicants, they require less staff time than personal interviews, while providing more support to the applicant than the self-administered form.

Some types of baseline data can be obtained from administrative records. For example, in experiments involving welfare recipients, baseline values of welfare benefits are probably best collected from administrative records, because those records constitute the official record of the amounts paid and are probably quite accurate. The records are also quite inexpensive to access, because they are virtually always automated. For other background information, however, such as demographic characteristics or work history, administrative systems are notoriously inaccurate, especially if program benefits do not depend directly on the data in question. Moreover, the definitions used for such variables may vary considerably from one local program to another, making uniformity of data problematic in multisite experiments.

When baseline data are collected directly from the sample member, it is essential that they be collected before random assignment, for several reasons. First, this requirement ensures that, by definition, baseline data are available for all individuals who are randomly assigned — *i.e.*, for the entire analysis sample. Attempts to collect baseline data after random assignment invariably result in missing data for a portion of the sample, either because the sample member cannot be located or because they are unwilling to cooperate with the interview. Control group members tend to be more difficult to interview for both of these reasons; thus, not only will data be missing, but it is likely to be missing differentially for the treatment and control groups. Such disparities between the two groups can bias the impact estimates.

A second reason for collecting baseline data prior to random assignment is that it assures that the sample member's responses will not be influenced by knowledge of his or her experimental assignment. This is particularly true of attitudinal data, since attitudes can potentially be affected by exclusion from the experimental program. But other types of data may be influenced as well; individuals who have been excluded from the program may simply not give their responses as much time and thought as those who have been accepted into the program. Control group members may, for example, provide less complete work histories than treatment group members; this would result in the appearance of a mismatch between the two groups in terms of prior work experience. "Correcting" this mismatch in the analysis would produce biased impact estimates.

Outcome Data

Anything that the treatment group does after random assignment is potentially affected by the experimental program — even if only through the individual's knowledge that he or she is eligible to participate in it. Therefore, the sole purpose of collecting data on the experimental sample after random assignment is to *provide the outcome data needed to estimate program impacts*.

A common error in designing follow-up data collection instruments is collecting data on variables that are viewed as useful in "explaining" the postprogram outcomes. As noted above in connection with baseline data, in an experiment it is not necessary to "explain" the outcomes; the control group provides all the information we need about what the outcomes would have been in the absence of the experimental program. In fact, since these "explanatory" variables may be affected by the experimental treatment, their inclusion in the model used to estimate impacts on other outcomes may bias the impact estimates by capturing part of the effect of the treatment. Thus, follow-up data collection should be focused on those variables that reflect the intended objectives of the program and/or any other potential consequences of program participation. (See the discussion of specification of outcomes in the second paper in this series.)

Outcome data can be collected either through surveys of the experimental sample or from administrative records. Surveys are typically conducted by telephone or, for those who cannot be contacted by telephone, in person. Administrative data may come from any of a variety of program record systems — *e.g.*, the eligibility and benefit payment systems of social programs, health insurance records, or the employer-reported wage records maintained by state Unemployment Insurance (UI) agencies for all UI–covered workers.

Where available, administrative records have several distinct advantages over surveys.[30] First, for some variables, they are likely to be more accurate. For example, health insurance records of the amount and cost of medical care consumed by an individual are likely to be more accurate than the individual's recall of those items. Second, administrative records are not subject to nonresponse, as surveys are. Thus, one can often obtain data for virtually 100 percent of the population of interest, rather than the 75 or 80 percent response rate that is typical in surveys. Third, it is often cheaper to collect computerized administrative data than to conduct surveys, especially for large samples, since the cost of accessing electronic files is largely independent of the number of records to be extracted. Finally, administrative data are often available for much longer retrospective periods than can be reliably collected in surveys, because of respondent recall error.

Administrative records also have important limitations, however. For many outcomes of interest, they are simply not available. In some cases, they may cover only part of the population of interest. For example, state AFDC payment systems will provide accurate records of benefits paid to families within the state, but no information on families who move to another state and receive benefits. Access to administrative records may also be restricted for reasons of confidentiality. Usually (although not always), this restriction can be removed by obtaining a signed release from the sample member; such releases are routinely collected as part of the baseline information form in many experiments.

[30] For a good discussion of the issues involved in using administrative data in one specific context--measuring the post-program earnings of JTPA participants--see Baj and Trott (1991) or Office of Technology Assessment (1994).

Even where administrative records are available, they often contain some data elements that are highly *inaccurate*. In general, the most reliable administrative data are those that record program benefits in quantitative terms (*e.g.*, welfare payments) or relatively objective personal characteristics that are central to eligibility for, or amount of, program benefits (*e.g.*, number and age of the children of welfare recipients). Descriptive data that are not essential for determining program eligibility or benefit amount (*e.g.*, educational attainment) are often collected only sporadically and/or carelessly by program staff, and may not be updated on a regular basis. Moreover, the content and definition of variables in administrative records may vary greatly from one local program to another, making it difficult to collect uniform data for all members of the sample.

Finally, for a variety of reasons, it is often difficult to obtain complete and accurate extracts of administrative records. Agency staff responsible for maintaining administrative data often give low priority to requests for such extracts, because they view them as a diversion from their programmatic mission. As a result, they may be careless in executing such requests or misunderstand what is wanted — for example, providing benefit records for *current* welfare recipients only, rather than all sample members who ever received welfare during the follow-up period. Mistakes may also arise simply out of lack of experience in producing such extracts. For reasons such as these, the National JTPA Study was able to collect complete and accurate data on UI–covered earnings in only 12 of 16 study sites and on AFDC and food stamps benefits in only four and two sites, respectively.

The great advantage of surveys is their ability to collect detailed data tailored to the analysis at hand. Their greatest limitations are cost, nonresponse, and respondents' limited knowledge of, or ability to recall, certain outcomes. A typical follow-up survey costs $100 to $200 per completed interview, depending on its length and the difficulty of locating the respondent population. For a sample of 10,000 respondents, the total cost of such a survey can easily exceed $1 million . Follow-up survey response rates seldom exceed 80 to 85 percent, and respondents may have great difficulty answering certain questions. For example, most people cannot reconstruct their expenditures or amounts of program benefits received with much precision and many cannot distinguish among the programs from which they or their families have received cash benefits (*e.g.*, AFDC, General Assistance, SSI, Social Security). These limitations notwithstanding, surveys are often the only available source of data on the experimental outcomes.

However the outcome data are collected, it is critically important to apply the *same* methods to the treatment and control groups, in order to avoid treatment–control differences in measured outcomes that reflect differences in data collection methods, rather than program impacts. Consider, for example, an experimental test of a health clinic in which utilization of medical care is an outcome of interest. One might be tempted to obtain medical utilization data for the treatment group from the clinic's administrative records, while using a survey to measure utilization among controls, who obtained their care from a wide range of providers in the community. Such a strategy would likely lead to differential underreporting between the two groups; this difference in reported utilization would then be mistakenly attributed to the impact of the program.[31]

A more subtle example of the confounding of experimental impacts with differences in data collection methods occurs when the records of the experimental program are used to update the contact information obtained at baseline. Since the program can only provide updated information for the treatment group, this results in more current information, and therefore higher survey response rates, for the treatment group than for controls. If survey respondents differ systematically from nonrespondents in ways that affect their outcomes, such differential nonresponse may lead to biased impact estimates. The only safe rule is to *use only those sources of contact information that are available symmetrically for both treatment and control groups.*

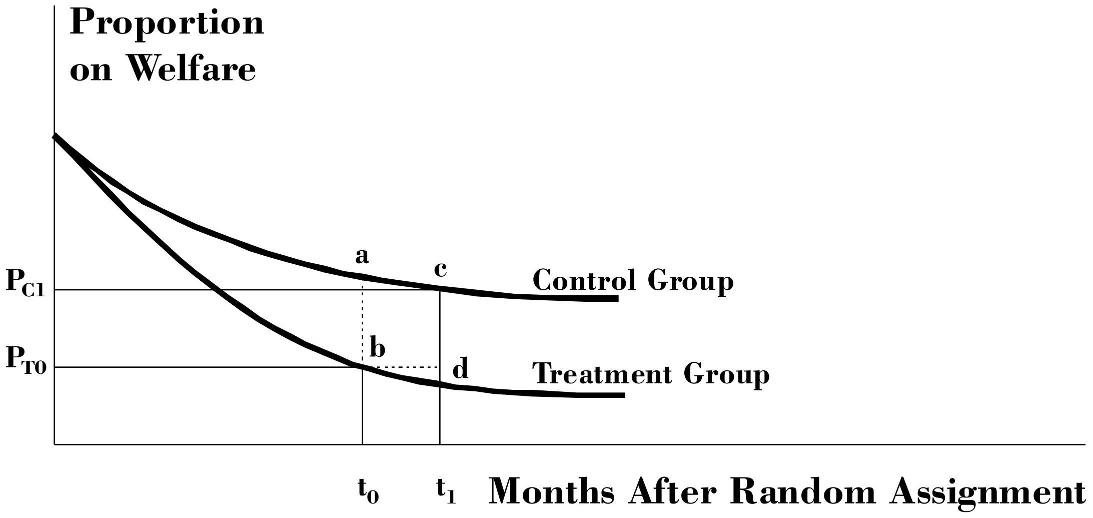
The timing of data collection must also be symmetric between the two groups. Exhibit 4 shows how differences in the timing of measurement can bias the impact estimates when there is a trend in the outcome of interest. In this hypothetical case, the proportion of the treatment group on welfare is measured at an earlier point relative to random assignment (t_0) than that of the control group (t_1). Because the rate of welfare receipt is falling in both groups, the treatment–control difference ($P_{T0} - P_{C1}$) measured this way understates the true impact of the program (the vertical distance between the two outcome lines) at either time t_0 (the distance ***ab***) or time t_1 (the distance ***cd***).

More generally, outcome data for all individuals in the experimental sample must be aligned according to time elapsed since random assignment, to avoid confounding the experimental impact with trends in the outcome vari-

[31] In this case, one could not even be sure of the direction of the bias. On the one hand, survey respondents are likely to underreport their utilization because of recall error, whereas the clinic's administrative records will accurately record all care received at the clinic. On the other hand, however, the clinic's administrative records will miss any care received from other sources.

Effect of Asymmetric Timing of Data Collection on the Impact Estimate

EXHIBIT 4



ables. This applies both to "snapshot" outcome measures, like that depicted in Exhibit 4, and to outcomes measured continuously over time (*e.g.*, monthly earnings).

Analysts sometimes attempt to key follow-up data collection to the point at which individuals leave the program — *e.g.*, conducting follow-up interviews six months after program exit. This approach is fundamentally flawed, because program exit is not defined for controls, who were excluded from the program. Treatment group members who leave the program early (or never enter it) are likely to differ systematically from those who leave it later — *e.g.*, they may be less motivated or, conversely, they may be more successful. There is no way to identify the control counterparts of these different self-selected groups, so that their follow-up interviews can be conducted at the same time.

Program Participation Data

In an experimental impact study, exclusion of the control group from the experimental program creates a service differential between the treatment group and the control group. Impacts are then estimated as the treatment–control differences in the outcomes of interest. Remarkably enough, no information on the actual treatment–control service differential is needed in order to estimate its impact. However, if the impact estimates are to be useful for policy, it is important to be able to describe the service differential that produced them so that the experimental program can be replicated on a larger scale. This information is also useful in interpreting the experimental results; for example, in some experiments these data have revealed that the reason the program had no impact was that the experimental program failed to create a significant service differential.[32]

For these purposes, it is useful to know the nature of the services or benefits provided and the duration of receipt of those services or benefits. For example, in a counseling program, we would want to know the number of counseling sessions, their timing and content, the credentials and approach of the counselor, and any services to which the individual was referred by the counselor. In a training program, we would want to know the dates of entry and exit, the type and number of hours of training received, the training curriculum, and whether the individual successfully completed the course.

In a program that does not displace services and benefits that would have been received from nonexperimental sources, the treatment–control service differential is simply the experimental services received by the treatment group. Thus, we need only collect information about these services in order to describe the treatment–control service differential. This information can usually be obtained from the administrative records of the organization providing the experimental services, although in some cases it will be necessary to set up special data collection systems to capture this information over the course of the experiment. For example, a counseling program might not normally keep track of the number, length, or content of individual coun-

[32] This can occur either because the program produces very little service or because the service it does produce simply displaces service from nonexperimental sources.

seling sessions. In such a case, it will be necessary to design a form on which counselors are asked to record these data as each session is held.

Collection of program participation data is much more difficult in cases where the experimental program can be expected to displace similar nonexperimental services. In such cases, in order to measure the *net* treatment–control service differential it is necessary to measure *both experimental and nonexperimental services* received by *both treatment and control group members.*

Several different approaches are available to do so. First, one can use the administrative data compiled by the experiment to measure the delivery of experimental services and attempt to access similar administrative data on nonexperimental services. For example, in an evaluation of a training program, one might search the records of other training providers in the community — *e.g.*, JTPA, the Employment Service, the JOBS program, etc. — for services to the experimental sample.[33]

If successful, this approach can yield relatively accurate data on service receipt. There are several problems with this approach, however. As noted above, programs may not routinely record detailed service information; in many programs, the administrative records may contain little more than the dates the individual entered and left the program. Moreover, if there are a large number of potential alternative service providers — as is often the case, especially in multisite studies — it can be extremely expensive and time-consuming to identify all the relevant providers, negotiate access to their records, and extract the data for the sample. Finally, if there is a large number of providers, there is a substantial risk that some will be missed, resulting in an underestimate of the nonexperimental services received by the control group and an overstatement of the treatment–control service differential.

Because of these problems, service receipt is frequently measured through follow-up interviews with the sample members themselves. The obvious problem with this method is that respondents may not know, or may not remember, detailed information about the services they received from various programs, especially if the recall period is more than a few months. Respondents may also have trouble identifying some of the services they have received. For example, in the National JTPA Study, sample members could not distinguish the subsidized employment they obtained through JTPA from regular jobs. Realistically, however, this may be the only feasible approach for measuring nonexperimental services when there is a large number of service providers in the study sites. In the National JTPA Study, for example, sample members reported receiving education and training services from over 400 schools and training institutions.

In cases where nonexperimental services must be measured through follow-up surveys, the experimenter must decide whether to measure *all* services through the survey or to rely on administrative data for experimental services and survey data for nonexperimental services. The latter approach uses the best available data for each type of service, but probably measures experimental services more accurately than nonexperimental services, thereby confounding the treatment–control service differential with measurement error. The former approach avoids this problem, at the cost of using less accurate data.

Which approach is preferred depends on the relative accuracy of the two types of data and the relative amounts of service received by the two groups. The tradeoffs involved can be illustrated by a simple example. Suppose that the treatment group received an average of 100 hours of experimental service (only) while controls averaged 40 hours of nonexperimental service (only), for a (true) service differential of 60 hours. Further suppose that administrative data from the experiment accurately represent all experimental services received, while survey respondents underreport services (both experimental and nonexperimental) by 20 percent. If services to the treatment group are measured with administrative data and services to controls with survey data, we will obtain estimates of 100 hours of service to the treatment group and 32 hours of service to the control group, for a service differential of 68 hours. If services to both groups are measured with survey data, our estimates will be 80 hours of service to the treatment group and 32 hours to the control group, for a service differential of 48 hours. The former overstates the service differential by eight hours; the latter understates it by 12 hours.

In this example, then, the mixed-mode approach provides a more accurate measure of the true service differential. However, if the rate of underreporting on the survey were lower, or the treatment–control difference in service receipt smaller, using the survey to measure services to both groups might well yield the more accurate estimate.

A possible refinement to the mixed-mode approach is to compute an estimate of the underreporting rate in the survey, based on the ratio of treatment group service receipt reported in the administrative data to that reported in the

[33] Note that in both cases, we would attempt to identify services to both the treatment group and the control group.

survey. This estimate can then be used to correct the controls' survey data for underreporting.

It should be noted that, in practice, it is often impossible for survey respondents to distinguish experimental services from nonexperimental services. In that situation, in order to use administrative data from the experiment to measure services one must assume that the treatment group received *only* experimental services and the control group received *only* nonexperimental services. The reasonableness of these assumptions will depend heavily on the nature of the services and the institutional setting of the experiment.

Cost Data

The great contribution of social experiments is to provide unbiased estimates of program impact. To be useful for policy, however, those estimates must be combined with information about the cost of the program. That is, policy makers must consider whether the program's impacts are sufficient to justify its costs. In a subsequent paper in this series, we will describe how such benefit–cost analyses are conducted. Here, we focus on the issues involved in collecting data on program costs.

The cost that must be measured is the cost of the treatment–control difference in services or benefits that produced the estimated impacts of the program — *i.e.*, the *net* or *incremental* cost of the experimental program. If the experimental program does not displace any nonexperimental services or benefits, this cost is simply the cost of the experimental program. If the experimental program bears the full cost of the services or benefits it provides, this cost can usually be measured with data from the administrative records of the experiment.

As with the measurement of service receipt, if the experiment displaces nonexperimental services, measurement of the net cost of the experimental program becomes more complex. In that case, we wish to measure:

$$C = c_t S_t - c_c S_c$$

where:

C = net cost of the experimental program per treatment group member

c_t = cost per unit of service received by the treatment group

S_t = average number of units of service received by treatment group members

c_c = cost per unit of service received by controls

S_c = average number of units of service received by controls

S_t and S_c are the treatment and control service levels whose measurement was discussed in the previous section. The problem here is to measure c_t and c_c, the unit costs of services received by the typical treatment and control group members. The fact that treatment and control group members may, in general, receive different types and intensities of service means that service costs must be measured separately for the two groups.

If treatment group members receive only experimental services and the experiment bears the full cost of those services, then $c_t S_t$, the average cost per treatment group member, can be calculated simply by dividing the total cost of the experimental program by the total number of treatment group members. If either of these conditions is not satisfied, however, the cost of services received by treatment group members must be estimated with data from sources other than the experiment. In any case, cost data must be collected from sources other than the experiment in order to estimate c_c, the unit cost of services to controls. How this is done will depend heavily on the specific services involved. Rather than attempting to prescribe any general guidelines for this task, we illustrate some of the possible sources of cost data by describing the approach taken in one experimental evaluation, the National JTPA Study.

In that study, receipt of employment and training services (S_i in the preceding equation) was measured through a combination of JTPA administrative data and a follow-up survey of the experimental sample.[34] JTPA administrative data were used to measure days of unpaid work experience and on-the-job training in subsidized jobs, on the grounds that these services were provided only by JTPA and that JTPA bore their full cost.[35] For those services that were readily available in the community from sources other than JTPA, survey data were used to measure the number of days or hours of service received by both treatment and control group members. This included such services as classroom training in vocational skills, basic education, and job search assistance.

The unit costs of these various services (c_i in the equation) were estimated on the basis of data from several sources. The costs per day of providing job search assistance, unpaid work experience, and on-the-job training in subsidized jobs were obtained by dividing total JTPA expenditures on each of these services by the total number of days of each

[34] See Orr et al. (1996), Appendix B, for a detailed description of the procedures used to estimate costs in the National JTPA Study.

[35] It would have been virtually impossible to measure these services through the follow-up survey in any case, because respondents could not reliably distinguish these types of employment from regular jobs.

provided to JTPA enrollees, as measured in JTPA administrative data. The cost of job search assistance estimated in this way was applied to assistance received from both JTPA and non–JTPA sources, as measured in the follow-up survey.

JTPA administrative data were not used to estimate the cost of classroom training in vocational skills and basic education, primarily because JTPA frequently does not pay the full cost of these services. JTPA often obtains such services from public institutions, such as community colleges and high schools, that receive substantial tax subsidies. Therefore, in the follow-up survey, respondents were asked to identify the specific institution at which they received classroom training, and data on the costs of the institutions named were obtained from other sources.

Hourly costs of instruction for each of the public high schools or colleges named in the survey were computed on the basis of institution-specific data compiled by the U.S. Department of Education. To estimate the cost of training at private schools and training institutions, which were not included in the Department of Education data, a telephone survey was conducted. In this survey, each of the private institutions named by follow-up survey respondents was contacted and asked their tuition rate. Since private schools receive no tax subsidy, their tuition can be treated as equal to the full social cost of the training they provide. The unit costs derived from these sources were then multiplied by the number of hours of classroom training reported on the follow-up survey to compute a total cost of training for each sample member.

References

Baj, John, and Charles E. Trott. 1991. *A Feasibility Study of the Use of Unemployment Insurance Wage-Record Data as an Evaluation Tool for JTPA.* Research Report No. 90-02. Washington, D.C.: National Commission for Employment Policy.

Boruch, Robert F., Michael Dennis, and Kim Carter-Greer. 1988. "Lessons from the Rockefeller Foundation's Experiments on the Minority Female Single Parent Program." *Evaluation Review, Vol. 12, No. 4.* Newbury Park, California: Sage Publications.

Doolittle, Fred, and Linda Traeger. 1990. *Implementing the National JTPA Study.* New York: Manpower Demonstration Research Corporation.

Office of Technology Assessment. 1994. *Wage Record Information Systems, OTA-BP-EHR-127.* Washington, D.C.: U.S. Congress.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study.* Washington, D.C.: Urban Institute Press.

Puma, Michael J., Janet DiPietro, Jeanne Rosenthal, David Connell, David Judkins, and Mary Kay Fox. 1991. *Study of the Impact of WIC on the Growth and Development of Children. Field Test: Feasibility Assessment.* Final Report: Volume I. Cambridge, Mass.: Abt Associates Inc.